GRNN: Generative Regression Neural Network—A Data Leakage Attack for Federated Learning

Year: 2021 | Citations: 119 | Authors: Hanchi Ren, Jingjing Deng, Xianghua Xie

---

Abstract

Data privacy has become an increasingly important issue in Machine Learning (ML), where many approaches have been developed to tackle this challenge, e.g., cryptography (Homomorphic Encryption (HE), Differential Privacy (DP)) and collaborative training (Secure Multi-Party Computation (MPC), Distributed Learning, and Federated Learning (FL)). These techniques have a particular focus on data encryption or secure local computation. They transfer the intermediate information to the third party to compute the final result. Gradient exchanging is commonly considered to be a secure way of training a robust model collaboratively in Deep Learning (DL). However, recent researches have demonstrated that sensitive information can be recovered from the shared gradient. Generative Adversarial Network (GAN), in particular, has shown to be effective in recovering such information. However, GAN based techniques require additional information, such as class labels that are generally unavailable for privacy-preserved learning. In this article, we show that, in the FL system, image-based privacy data can be easily recovered in full from the shared gradient only via our proposed Generative Regression Neural Network (GRNN). We formulate the attack to be a regression problem and optimize two branches of the generative model by minimizing the distance between gradients. We evaluate our method on several image classification tasks. The results illustrate that our proposed GRNN outperforms state-of-the-art methods with better stability, stronger robustness, and higher accuracy. It also has no convergence requirement to the global FL model. Moreover, we demonstrate information leakage using face re-identification. Some defense strategies are also discussed in this work.